A FŐKOMPONENS ANALIZIS HELYETT: KANONIKUS KOMPONENS ELEMZÉS ALKALMAZÁSA ADATREDUKCIÓRA A KIVÁLTOTT POTENCIÁLOK OSZTÁLYOZÁSÁBAN

Vitrai József, Czobor Pál, Simon Gábor, Varga László, Marosfi Sándor

Semmelweis Orvostudományi Egyetem Psychiátriai Klinika és Számitástechnikai Csoport, Budapesti Hőerőmü Vállalat

A kutatók, hipotézisük igazolására gyakran alkotnak olyan kisérleti modellt, amelyben vizsgálati objektumok csoportjait ugy próbálják elkülöniteni, hogy egyidejüleg megkeresik azokat a paramétereket, amelyek a csoportok eltérő sajátosságait jól tükrözik.

Elméleti és gyakorlati szempontok miatt egy másik probléma megoldása is ugyanakkor gyakran kivánatos, nevezetesen, a diszkriminálás-hoz szükséges változók számának redukciója.

Elvileg két lehetőség adódik az előzőekben megfogalmazott kettős cél eléréséhez: még a diszkriminális *előtt* csökkentjük az eseteket reprezentáló változók számát, vagy csupán eleve olyan mesterséges változókat hozunk létre, melyek a szétválasztás szempontjából jelentősek, és ily módon redukáljuk az eredeti dimenziók számát.

Ismereteink szerint a kiváltott potenciálok elemzésére az első lehetőséget ezideig alapvetően ugy próbálták kihasználni, hogy vagy valamilyen *a priori* elv alapján kis számu kiváltott potenciál paramétert definiáltak /pl.: csucsamplitudó-, latencia, szinuszoid vagy Gauss-görbékhez való hasonlóság/, vagy néhány, az adatok belső sajátosságaiból adódó, *a posteriori* változót állitottak elő főkomponens analizissel /PCA/, majd az igy jellemzett kiváltott potenciálokat diszkriminancia analizisnek /DA/ vetették alá [1].

Második lehetőségként eddig az eredeti adatok lépésenkénti diszkriminancia analizisének /SWDA/ lehetősége merült fel [2].

Az elválasztás optimalitása szempontjából azonban e módszerekkel kapcsolatban komoly problémák vethetők fel. Nyilvánvaló ugyanis, hogy mind az "önkényes" kiváltott potenciál paraméterek használata, mind az SWDA alkalmazása információ veszteséghez vezethet, amelynek szignifikanciáját igen nehéz felbecsülni [3]. A PCA-nak az elválasztásban felmerülő problémáit a kiváltott potenciálok elemzésében mindezideig nem tisztázták megfelelően. Az elmélet alapján ugyanis a PCA olyan adattömöritő eljárás, amely a diszkriminálás szempontjából *nem optimális*, mivel optimális tulajdonságai csak a teljes variancia leirása esetén mutatkoznak meg [4]. Ugyanakkor a diszkriminálás csupán a csoportok közötti, azaz between-varianciát hasznositja.

Ezen előadás anyagát egy olyan vizsgálat tapasztalatai adják, amelyben a between-varianciát optimálisan - a lehető legjobb adatredukció megvalósitásával - leiró módszert, a kanonikus komponens elemzést tanulmányoztuk.

Módszer

A kanonikus komponens elemzés /CCA/ elve

Az általunk kipróbált CCA módszer lényegében a kanonikus korreláció felhasználása a vizsgálati populációt alkotó *esetcsoportok* eltérő vonásainak optimális jellemzésére. Optimális a következő értelemben: az eljárás azt a legfeljebb g-1 /g: a csoportok száma/ mesterséges változót keresi meg, amelyek segitségével a csoportok a lehető legjobban elválaszthatók.

Először magáról a kanonikus korrelációról néhány szót: Ha a populációt jellemző változókat valamilyen célszerü szempontból két csoportra bontjuk, és e két *változó-csoportot* oly módon transzformáljuk, hogy mérhető legyen sztochasztikus kapcsolatuk, kanonikus korrelációt hajtunk végre. A két változócsoport közti összefüggés erősségét a kanonikus korrelációs koefficiens méri. Az átalakitás eredményeképpen a populációt alkotó *esetek*, a két transzformált változó-készlet terében hasonló alakzatban fognak elhelyezkedni. A hasonlóság mértéke a kanonikus korrelációtól függ. Másképpen, a transzformált változók, vagyis a kanonikus változók a populáció azon varianciáját képesek tükrözni, melyet a két változó-csoport *egyaránt* leir.

A kanonikus korrelációt ugy tudjuk *eset-csoportok* eltérő tulajdonságainak jellemzésére felhasználni, hogy u.n. hipotézis-változókat vezetünk be a populáció csoportokra bontásához. Ez tulajdonképpen azt jelenti, hogy mindazon eset, amely egy adott csoportba tartozik, azonos, a csoportra jellemző hipotézis-változó értéket kap. Ha most az eredeti változókra és a hipotézis-változókra kanonikus korrelációt alkalmazunk, az igy nyert kanonikus változók azt a varianciát irják le optimálisan, amelyet mind az eredeti-változók, mind a hipotézis-változók egyaránt jellemeznek. Ez pedig nem más, mint a csoportok közti, azaz between variancia!

Mivel a fenti elképzelések helyessége elméletileg igazolható, a CCA-t követő DA jobb eredményt kell, hogy szolgáltasson, mint a PCA utáni. Ennek gyakorlati kipróbálását a következőkben ismertetendő kiváltott potenciál vizsgálatokban végeztük el.

Vizsgálati anyag

E vizsgálatban 6 személy vett részt. 24 óra alatt 14, ill. két személy esetében 16 ülésben, ülésenként 4-féle inger segitségével egy-egy ingerhelyzetben 64 egyedi potenciálból átlagolt kiváltott potenciálokat nyertünk. Igy 4 személy esetében 56, két személynél 80 átlagolt kiváltott potenciálból állt a vizsgálati populáció. Egy-egy kiváltott potenciál közel 500 ms-os szakaszát 93 adat reprezentálta. Az amplitudót 5 ms-onként mintavételeztük.

Adatfeldolgozás

Az egyes vizsgálati személyek kiváltott potenciál-populációjára külön-külön PCA-t /kovarianciamátrixból számolva, rotálás nélkül/, majd ennek eredményeképpen nyert főkomponensértékekre, személyenként SWDA-t hajtottunk végre az eltérő ingerszituációkban kapott kiváltott potenciál-csoportok elkülönitésére.

A CCA megvalósitása során minden csoporthoz egy-egy hipotézisváltozót rendeltünk, igy a 93 változó /4 vagy 5/ további, az egy-egy adott csoportba való tartozást jellemző változókkal bővült. Egy adott kiváltott potenciál hipotézis-változóinak értékei közül az volt 1, amely ahhoz a csoporthoz volt rendelve, melybe ez az eset tartozott. E kiváltott potenciál többi hipotézis-változójának értékét 0-nak választottuk. Az igy előkészitett kiváltott potenciálokat az előzőekben megadott elveknek megfelelően kanonikus korrelációnak vetettük alá.

I. TABLE

	PCA + SWDA				CCA + SWDA				
	n_{EP}	mean correct classification jackkn. %	n_{comp}	n_{EP} var.	n_{grops}	mean correct classification jackkn. %	n_{comp}	$n_{var.}$	diff. %
D.J.	56	96.4	4	55	4	98.2	3	36	1.8
V.J.	54	92.6	4	53	4	100.0	3	33	7.4
J.L.	56	92.9	6	55	4	96.4	3	34	3.5
B.É.	54	89.1	5	53	4	92.7	3	32	3.6
C.P.	75	74.7	3	74	5	96.0	4	36	21.3
K.M.	80	75.0	5	79	5	92.5	4	34	17.5
								mean	9.18
								S.D.	8.21

A kanonikus változók értékein szintén SWDA-t számoltattunk. Mindhárom számitógépes módszert a Californiai Egyetem BMDP programcsomagjában [5] található programokkal hajtottuk végre.

Eredmények

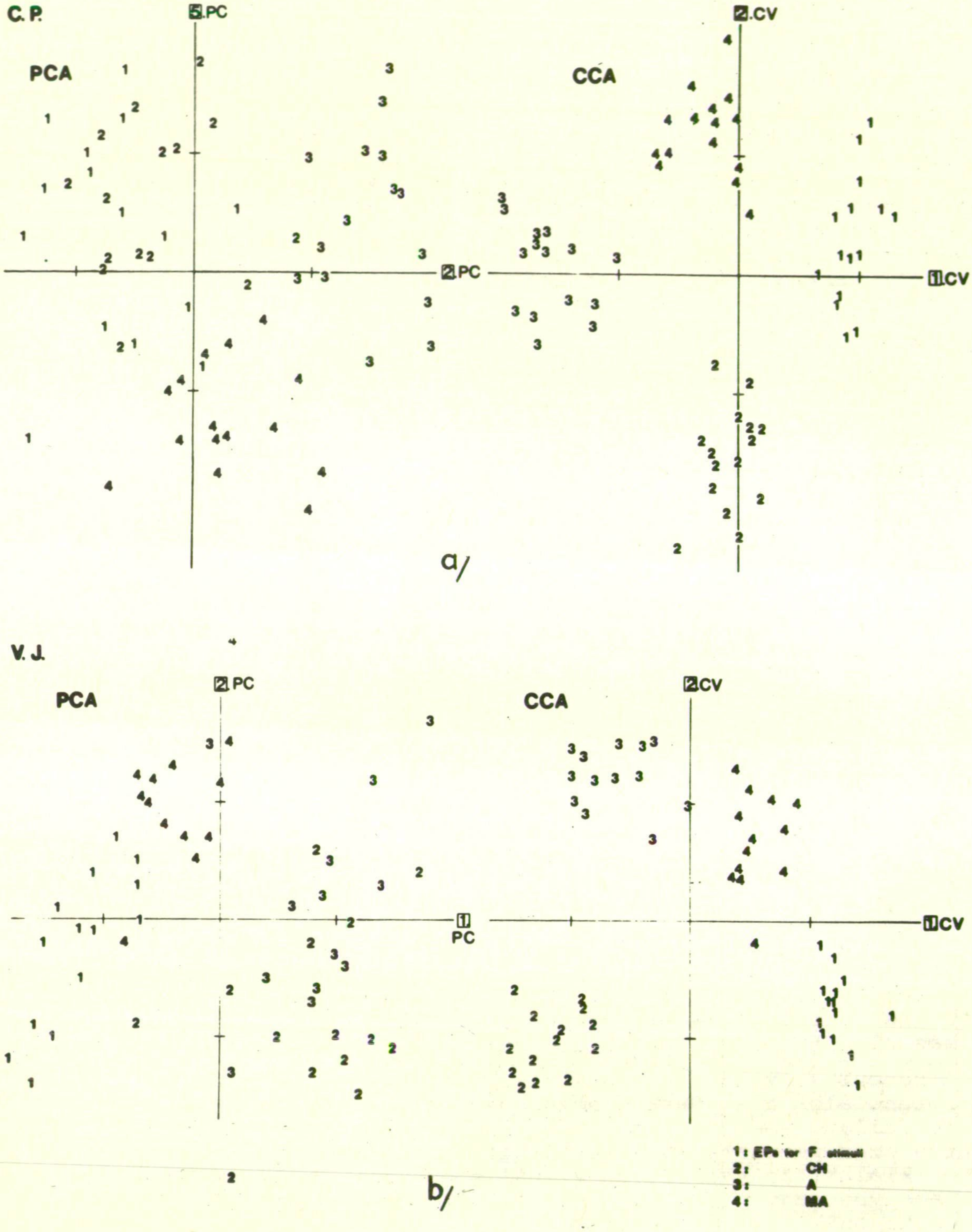
Annak ellenére, hogy a PCA-t követő SWDA is meglehetősen jó elválasztást nyujtott, a CCA által transzformált kiváltott potenciálokra végrehajtott SWDA képes volt további javulást szolgáltatni. A kétféle eljárással nyert diszkriminálás eredményeinek összehasonlitását teszi lehetővé az I. Táblázat. Megjegyezzük, hogy mivel a CCA-t követő SWDA mindegyik kanonikus változót bevonta a diszkriminálásba, tulajdonképpen egyszerü DA történt. A CCA az eredeti 93 kiváltott potenciál változónak csupán egy részét használta fel a kanonikus változók kialakitásához /$n_{EP\ var}$ az I. Táblázatban/. A többi az elválasztás szempontjából rendundásnak bizonyult.

Mint az előzőekben megfogalmazott, a CCA-ra vonatkozó optimalitásból következik, az összes lineáris /ortogonális/ transzformációk közül a CCA nyujtja a legjobb elválasztást egy legfeljebb g-1 dimenziós térben. Azonban az I. Táblázatból az is kitünik, hogy a PCA-t követő SWDA a legjobb elválasztást nyujtó, g-1-nél több változót felhasználva sem tud ott a CCA-t követő SWDA-nál nagyobb helyes besorolási arányt elérni. Két-két mesterséges változó alapján, azaz két dimenzióban összehasonlitva a PCA és a CCA hatékonyságának különbsége hasonlóan impresszionáló /1. ábra/.

Érdemes külön tárgyalni az F1 és F2 ingerlési szituációban nyert kiváltott potenciálok elválasztásában a kétféle eljárás eredményességét. Az F2 inger ugyanis az F1 ingernek az ülés végén 10-15 perc mulva történő megismétlése volt. Igy természetes, hogy a két kiváltott potenciál-csoport különbözősége nem mondható tul markánsnak /2. ábra/. A két eljárás teljesitményének különbözősége e két nehezebben elválasztható kiváltott potenciál-csoport esetén is kifejezett marad /II. Táblázat/.

A kanonikus változók tipikus "különbség potenciálokként" értelmezhetők. Ezek meghatározott összegzésével /a kanonikus változó értéke alapján/ mindegyik kiváltott potenciálra előállitható a kizárólag a saját csoportjára jellemző alakzat. Ez egy olyan szürésnek fogható fel, amelynek során az egyedi kiváltott potenciálból kiemeljük a csoportjára jellemző alakzatot, és ezt a kanonikus változók összegeként adjuk meg. Hangsulyozni kell azonban, hogy ez a "csoportjára jellemző alakzat" nem a csoport inherens sajátossága, hanem a többi csoporttól való eltérésével meghatározott, külső tulajdonság. Ugyanaz a csoport más és más csoprttal együtt elemezve, másként térhet el a többitől, azaz más alakot mutató kanonikus változóval irható le.

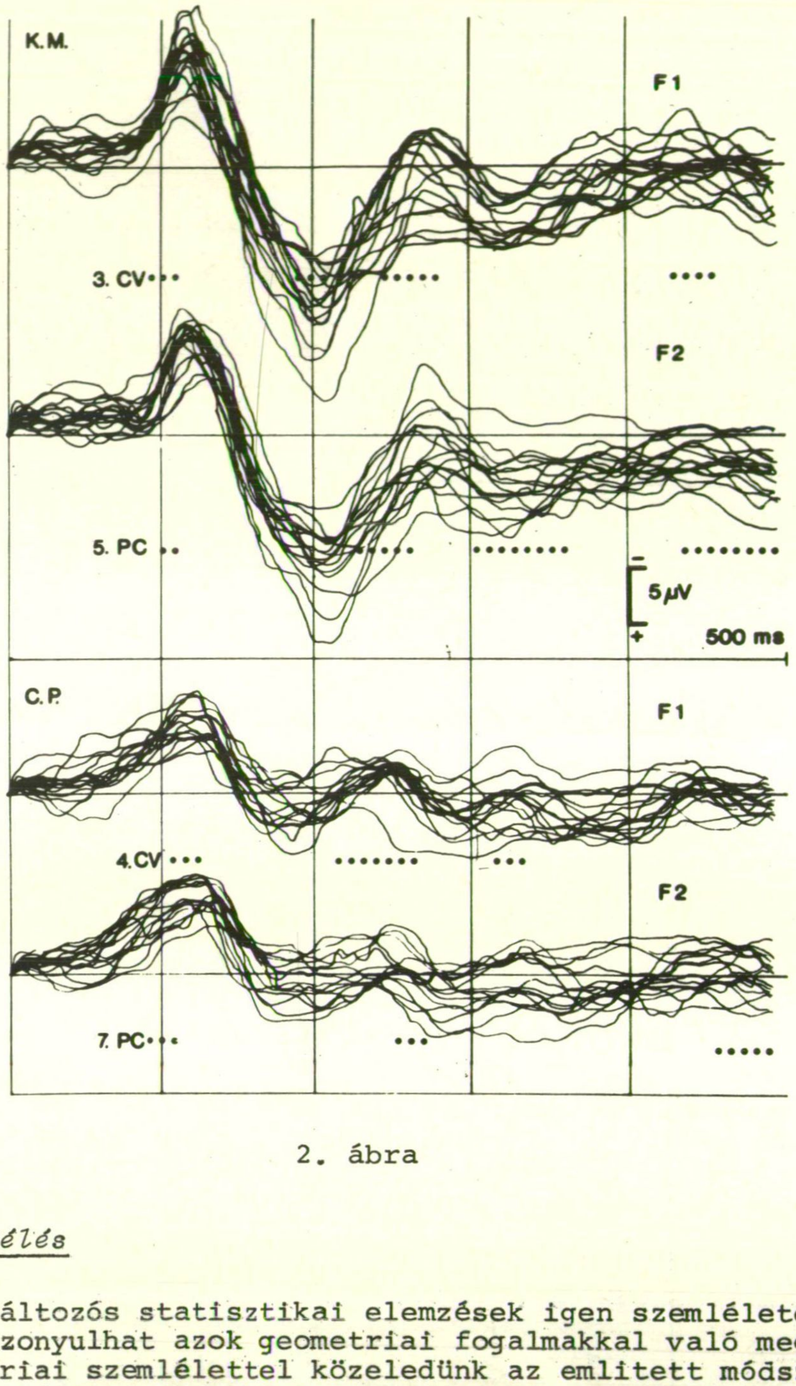
A kanonikus változók alaki kiértékelése hozzásegithet a kiváltott potenciálok alakjának elemzéséhez. Az a kiváltott potenciál időintervallum, amelyben egy kanonikus változó nagy értéket vesz fel, jelentős szerepet játszik a kiváltott potenciál-csoportok különbözőségének kialakitásában. Ilyen vizuális elemzéshez használható a 3. ábra, amelyben egy vizsgálati személy átlagolt kiváltott potenciáljait és kanonikus komponenseit tüntettük fel. Az ily módon nyert információk alapján a későbbiekben, a gyakorlati felhasználás során, már a CCA-nál jóval egyszerübb, esetleg "kézi" méréssel is végezhető kiváltott potenciál elemzés.



1. ábra

II. TABLE

		PCA + SWDA			CCA + SWDA		
	n_{F1+F2}	mean correct classification jackkn. %	n_{comp}	significance of F1-F2 distance by U-statistic p	mean correct classification jackkn. %	n_{comp}	significance of F1-F2 distance by U-statistic p
C.P.	15+15	66.7	2	10^{-1}	90.0	4	10^{-13}
K.M.	16+16	78.1	4	10^{-3}	87.5	3	10^{-12}

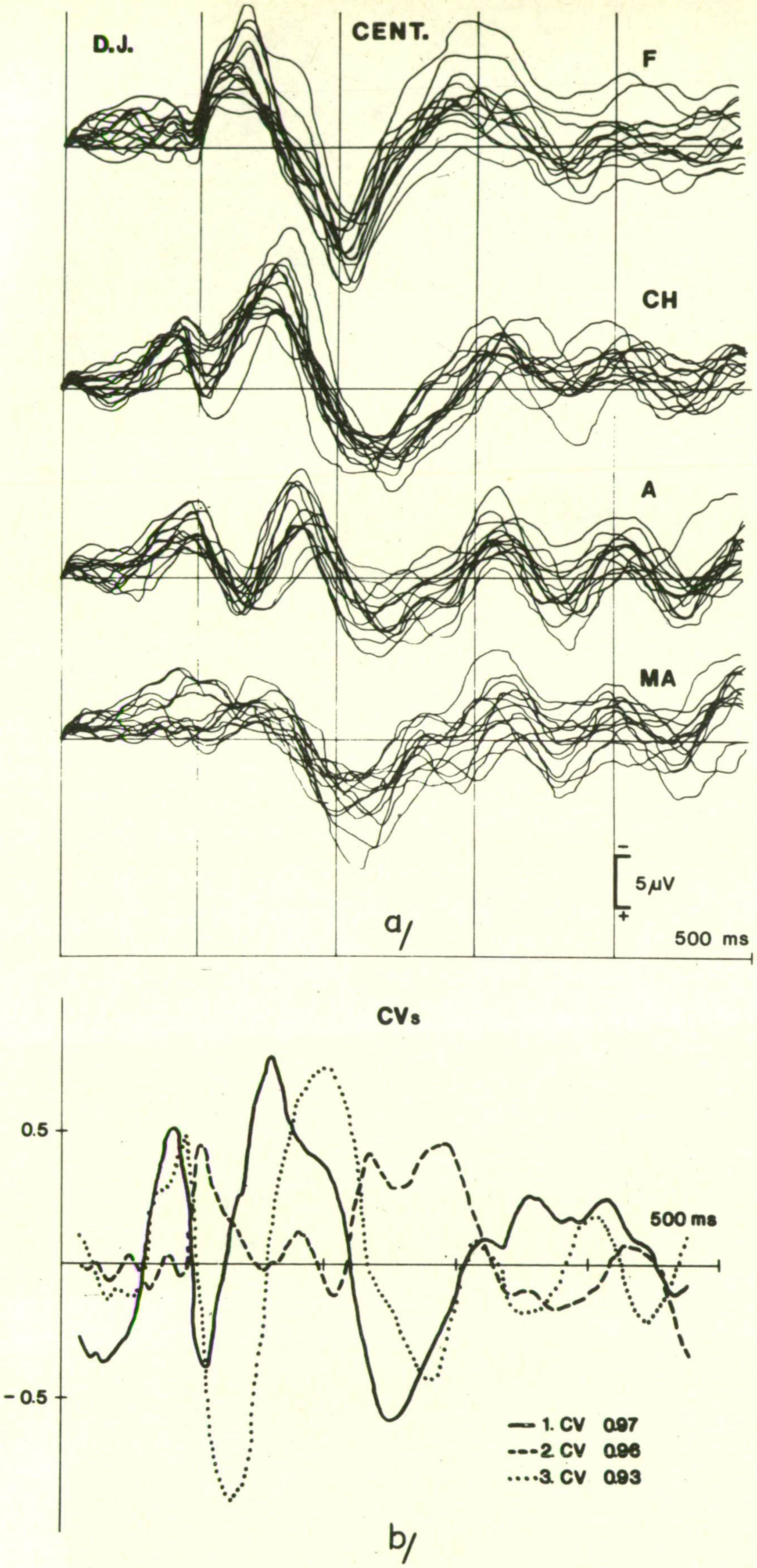


2. ábra

Megbeszélés

A többváltozós statisztikai elemzések igen szemléletes megközelitésének bizonyulhat azok geometriai fogalmakkal való megragadása. Ha ilyen geometriai szemlélettel közeledünk az emlitett módszerekhez, könnyen megvilágithatók a közöttük fennálló viszonyok.

Eszerint a PCA-t ugy lehet interpretálni, hogy ez az eljárás olyan mesterséges változókat, főkomponenseket számit ki, melyek segitségével legtöbbször az eredeti térnél kisebb dimenziószámu altérbe vetithetjük a reprezentált eseteket. A kisebb dimenziószám ellenére a teljes variancia nagy hüséggel leiródik ebben az altérben is. A transzformált adatokra végrehajtott DA ebben az altérben keres olyan elválasztó sikokat, amelyek egy-egy csoportot a többitől optimálisan elkülönitenek. Ezen sikok szám g. Mivel tehát a PCA által megvalósi-


D.J.
CENT.
F
CH
A
MA
5 μV
500 ms
a/
CVs
0.5
- 0.5
500 ms
1. CV 0.97
2. CV 0.96
3. CV 0.93
b/

3. ábra

tott transzformáció a teljes, az utána következő DA viszont a between variancia leirására nézve optimális, a két módszer együttes használata elvileg *sem* nyujthat az adatredukcióra és az elválasztásra nézve optimális megoldást.

A CCA-hoz is rendelhető geometriai jellegü leirásmód. Ebben a megközelitésben az esetek csoportjai eredetileg két térben helyezhetők el: a hipotézis-, ill. az eredeti változók terében. A CCA e két térre olyan transzformációt hajt végre, melynek eredményeképpen az esetek mindkét transzformált altérben hasonló konfigurációban fognak elhelyezkedni. Az általunk megválasztott hipotézis-változók inherens tulajdonsága miatt, ezek ortogonális transzformációjával létrehozott altérben az esetek adekvát módon, azaz a priori besorolásuknak megfelelően fognak csoportosulni. Mivel az eredeti változók generálta tér transzformációja olyan alteret eredményez, amelyben az esetek hasonló alakzatban találhatók mint a hipotézis-változók terében, igy ebben a transzformált térben is az a priori besorolásuknak megfelelően fognak az esetek csoportosulni. A két tér dimenziószáma megegyezik: g-1. Belátható, hogy a hipotézis-változók - természetüknél fogva - a csoporton belüli varianciát, a within varianciát egyáltalán nem tükrözik, igy a CCA a csoportok közötti különbség leirására optimális.

Ezek az előnyös tulajdonságok csak bizonyos feltételek teljesülése mellett mutatkoznak meg. Igy a csoportok kovarianciájának hasonlónak kell lennie, ugyanakkor további követelmény az is, hogy a változók *együttes* eloszlása normális legyen. Ezek a feltételek tulajdonképpen a *lineáris diszkriminálás* általános feltételei, és nem csupán e módszerre jelentenek megszoritást. .

Különösen előnyös tulajdonsága a CCA-nak, hogy az alapadatok lineáris transzformációjára érzéketlen. Igy alkalmazásakor választhatunk tetszőleges egységet, bázist, akár standardizálatlan adatokkal is dolgozhatunk. Ez tulajdonképpen abból következik, hogy a kanonikus korreláció a korreláció általánositásának tekinthető.

Fontos kiemelni, hogy a CCA és a DA lényegében ugyanannak a problémának kétféle megközelitése [6, 7].

Ugyanakkor a sajátértékprobléma megoldásán alapuló módszerek számitógépes megvalósitása nagy számitásigényü, tehát meglehetősen költséges futtatni azokat. Igy felmerülhet az a kérdés, hogy ha a CCA is és a DA is ugyanazon probléma egyenértékü megoldásai, érdemes lehet az "olcsóbb" DA-t használni. Abban az esetben, ha a csoportok eltérő vonásainak, vagyis különbözőségüknek jellemzése *is* célunk, a CCA-t, azonban ha maga az elválasztás csupán a cél, inkább a DA-t érdemes választani.

A CCA alkalmazása a PCA helyett mindazokban a vizsgálatokban különösen előnyös, amelyekben az elkülönitendő csoportokon belüli variancia, a within, a between-hez képest viszonylag nagy. /Ebben az esetben a PCA az optimálistól jelentősen eltérő eredményt produkálhat./

Végül felhivjuk a figyelmet arra, hogy a hipotézis-változók célszerü megválasztásával a CCA nemcsak diszkriminálásra, hanem másfajta feladat megoldására is alkalmas lehet. A CCA értelmezést ugy lehet általánositani, hogy az olyan módszer, amely egy vagy több előre megadott szempont alapján jellemzi, azaz bontja komponensekre a vizsgált jeleket. Egyszerü példának hozható fel egy olyan kisérleti modell lehetősége, amelyben az input-paraméterek változtatásának az outputon mutatkozó hatását kivánjuk elemezni. Ekkor a CCA-ban a hipotézis-változók az input-paraméterekkel egyeznének meg, és az eredményül nyert komponensek jellemeznék az input változásának hatását az outputon, ha lineáris összefüggéseket feltételeznek.

Referenciák

[1] McCallum, W.C. and Curry, S.H.: Late slow wave components of auditory evoked potentials: their cognitive significance and interaction. Electroenceph. clin. Neurophysiol., 1981, 51: 123-137.

[2] Horst, R.L. and Donchin, E.: Beyond averaging. II. Single-trial classification of exogenous event-related potentials using stepwise discriminant analysis. Electroenceph. clin. Neurophysiol., 1980, 48: 113-126.

[3] Daruna, J.H. and Karrer, R.: On the validation of discriminant functions: an empirical analysis using event-related potentials. Psychophysiology, 1981, 18: 82-87.

[4] Donchin, E. and Heffley, E.: Multivariate analysis of event--related potential data: A tutorial review. In: D.A. Otto /Ed./, Multidisciplinary Perspectives in Event-Related Brain Potential Research. EPA-600/9-77-043, U.S. Government Printing Office, Washington, D.C., 1978: 555-572.

[5] Dixon, W.J. /Ed./: BMD-Biomedical Computer Programs. University of California Press, Los Angeles, Calif., 1970.

[6] Lachenbruch, P.A.: Discriminant Analysis. Hafner, New York, 1975: 63-72.

[7] Jennrich, R.I.: Stepwise discriminant analysis. In: K. Enslein, A. Ralston, H.S. Wilf /Eds./, Statistical Methods for Digital Computers, vol. 3. John Wiley and Sons, Inc., 1977: 76-95.